**DEPARTMENTOFCOMPUTERSCIENCEANDENGINEERING**

**MACHINELEARNINGLABLISTOFEXPERIMENTS**

|  |  |
| --- | --- |
| **S.No.** | **NameoftheExperiment** |
| 1. | The probability that it is Friday and that a student is absent is 3 %. Sincethereare5schooldaysinaweek,theprobabilitythatitisFridayis 20%.WhatistheprobabilitythatastudentisabsentgiventhattodayisFriday?  ApplyBaye’sruleinpythonto gettheresult.(Ans:15%) |
| 2. | Extractthedatafromdatabaseusingpython |
| 3. | Implementk-nearestneighboursclassificationusingpython |
| 4. | Given the following data, which specify classifications for nine ombinationsof VAR1 and VAR2 predict a classification for a case where VAR1=0.906andVAR2=0.606,usingtheresultofk-meansclusteringwith3means(i.e.,3centroids)  VAR1 VAR2 CLASS1.713 1.586 0  0.180 1.786 1  0.353 1.240 1  0.940 1.566 0  1.486 0.759 1  1.266 1.106 0  1.540 0.419 1  0.459 1.799 1  0.773 0.186 1 |
| 5. | **Thefollowingtrainingexamplesmapdescriptionsofindividualsontohigh,mediumandlow** |

|  |  |
| --- | --- |
|  | credit-worthiness.  medium skiing design single twenties no ->highRiskhigh golf trading married fortiesyes->lowRiskow speedway transport married thirties yes ->medRiskmedium footballbanking single thirties yes ->lowRiskhigh flying media marriedfifties yes ->highRiskow footballsecurity single twenties no ->medRiskmedium golf media single thirties yes ->medRiskmedium golf transport married forties yes ->lowRiskhigh skiing banking single thirties yes ->highRiskow golf unemployedmarriedfortiesyes->highRisk  Input attributes are (from left to right) income, recreation, job, status, age-group, home-owner. Find the unconditional probability of `golf' and theconditionalprobabilityof`single'given`medRisk'inthe dataset? |
| 6. | Implementlinearregressionusingpython. |
| 7. | ImplementNaïveBayestheoremtoclassifytheEnglishtext |
| 8. | Implement an algorithm to demonstrate the significance of geneticalgorithm |

# DEPARTMENTOFCOMPUTERSCIENCEANDENGINEERINGMACHINELEARNINGLAB

**INTRODUCTIONTOLAB:**

Machine Learningis used anywhere from automatingmundane tasks to offeringintelligentinsights,industries in every sector try to benefit from it. You may already be using a device that utilizes it. Forexample, a wearable fitness tracker like Fitbit, or an intelligent home assistant like Google Home.Butthere aremuchmoreexamples ofMLinuse.

* **Prediction:**Machine learning can also be used in the prediction systems. Considering the loanexample, to compute the probability of a fault, the system will need to classify the available data ingroups.
* **Image recognition**:Machine learning can be used for face detection in an image as well. There is aseparate categoryforeachpersoninadatabaseofseveralpeople.
* **Speech Recognition:**It is the translation of spoken words into the text. It is used in voice searchesand more. Voice user interfaces include voice dialing, call routing, and appliance control. It canalsobeusedasimpledataentryandthepreparationofstructureddocuments.
* **Medicaldiagnoses:**MListrained to recognizecanceroustissues.
* **Financialindustry:**andtrading:companiesuseMLinfraudinvestigationsandcreditchecks.

## TypesofMachineLearning?

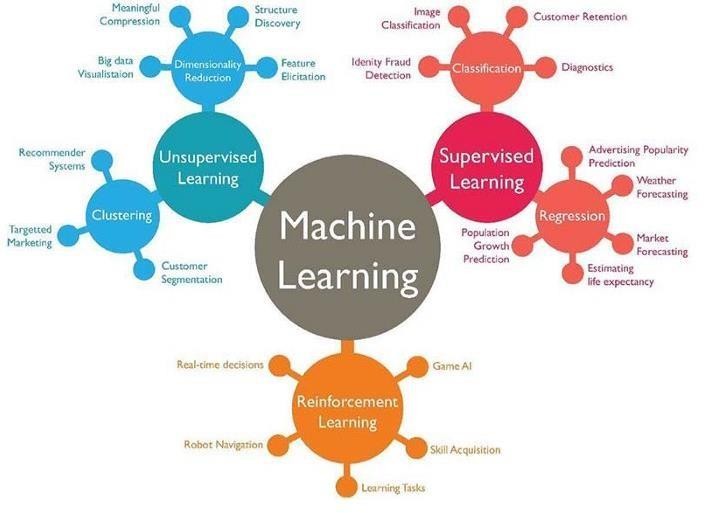
Machinelearningcanbeclassifiedinto3typesofalgorithms

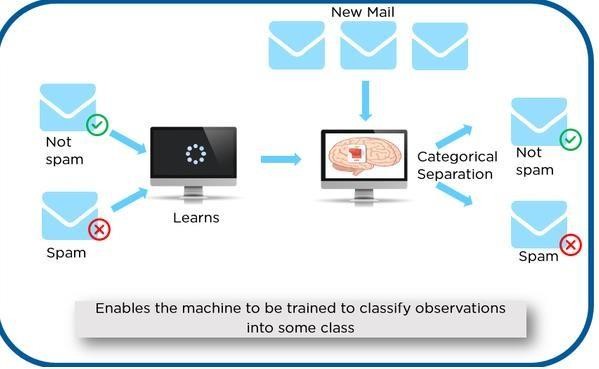
1. SupervisedLearning
2. UnsupervisedLearning
3. ReinforcementLearning

## OverviewofSupervisedLearningAlgorithm

InSupervisedlearning,anAI systemispresented withdatawhichislabeled,whichmeansthateachdatataggedwiththecorrectlabel.

Thegoalistoapproximatethemappingfunctionsowellthatwhenyouhavenewinputdata(x)thatyoucanpredicttheoutputvariables(Y)forthatdata.





As shown in the above example, we have initially taken some data and marked them as ‘Spam’ or ‘NotSpam’.Thislabeleddataisusedbythetraining supervisedmodel,thisdata isusedtotrainthemodel.

Once itis trained we can test our model by testing it with some test new mails and checking of the modelisabletopredicttherightoutput.

## TypesofSupervisedlearning

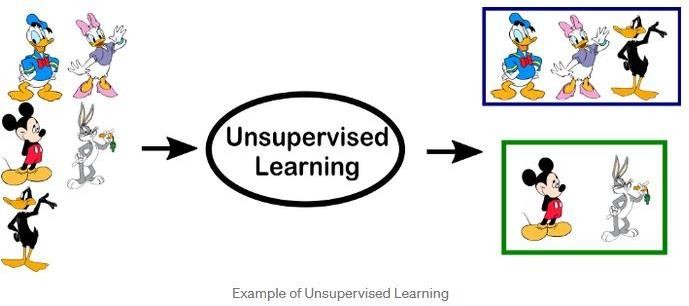
* **Classification**: A classification problem is when the output variable is a category, such as “red” or“blue” or“disease”and“nodisease”.
* **Regression**:Aregressionproblemiswhentheoutputvariableisarealvalue,suchas“dollars” or“weight”.

## OverviewofUnsupervisedLearningAlgorithm

In unsupervised learning, an AI system is presented with unlabeled, uncategorized data and the system’salgorithms act on the data without prior training. The output is dependent upon the coded algorithms.Subjectingasystemtounsupervisedlearningis onewayoftestingAI.

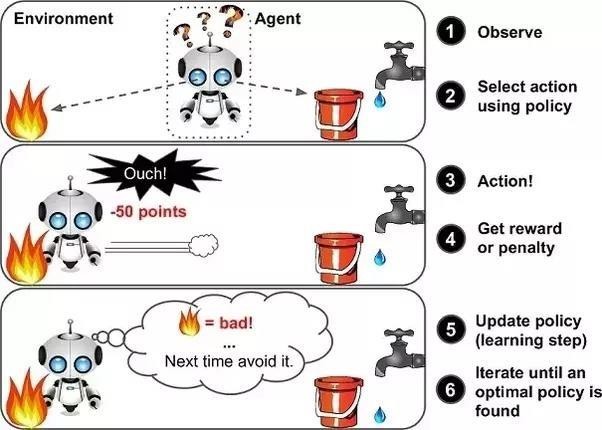
## TypesofUnsupervisedlearning:

* **Clustering**:Aclusteringproblemiswhereyouwanttodiscovertheinherentgroupingsinthedata,suchas groupingcustomers bypurchasingbehavior.
* **Association**:Anassociationrulelearningproblemiswhereyouwanttodiscoverrules thatdescribelargeportionsofyourdata,suchaspeoplethatbuyXalsotendtobuyY.



## OverviewofReinforcementLearning

A reinforcementlearning algorithm,or agent, learnsby interacting with its environment. The agentreceives rewardsby performing correctly and penaltiesforperformingincorrectly.The agentlearnswithout intervention from a human by maximizing its reward and minimizing its penalty. It is a type ofdynamicprogrammingthattrainsalgorithmsusingasystemofrewardandpunishment.



in the above example, we can see that the agent is given 2 options i.e. a path with water or a path with fire.A reinforcement algorithm works on reward a system i.e. if the agentuses the fire path then the rewardsare subtracted and agent tries to learn that it should avoid the fire path. If it had chosen the water path orthe safe path then some points would have been added to the reward points, the agent then would try tolearnwhatpathissafeandwhatpathisn’t.

It is basically leveraging the rewards obtained; the agent improves its environment knowledge to select thenextaction.

## PROGRAM1

**TheprobabilitythatitisFridayandthatastudentisabsentis3%.Sincethereare5schooldaysinaweek,theprobabilitythatitisFridayis20%.Whatistheprobabilitythatastudentisabsent given that today is Friday? Apply Baye’s rule in python to get the result. (Ans:15%)**

**AIM:**TofindtheprobabilitythatastudentisabsentgiventhattodayisFriday.

## DESCRIPTION:

Machine learning is a method of data analysis thatautomates analytical model building of data set.Using the implemented algorithms that iteratively learn from data, machinelearning allows computersto find hidden insights without being explicitly programmed where to look. Naivebayes algorithm isone of the most popular machines learning technique. In this article we will look how to implementNaivebayes algorithmusingpython.

Before someone can understand Bayes’ theorem, they need to know a couple of related concepts first,namely,theideaofConditionalProbability,andBayes’Rule.

Conditional Probability is just what is the probability that something will happen, given that somethingelse has alreadyhappened.

Let say we have a collection of people. Some of them are singers. They are either male or female. If weselect a random sample, what is the probability that this person is a male? whatis the probability thatthis person is a male and singer? Conditional Probability is the best option here. We can calculateprobabilitylike,

P(Singer&Male)=P(Male)xP(Singer/Male)

## WhatisBayesrule?

We can simply define Bayes rule like this. Let A1, A2, …, An be a set of mutually exclusive events thattogether form the sample space S. Let B be any event from the same sample space, such that P(B) > 0.Then,P(Ak|B)=P(Ak∩B) /P(A1∩ B )+P(A2∩B)+. ..+P(An∩ B )

## WhatisBayesclassifier?

NaiveBayesclassifiersareafamily of simpleprobabilisticclassifiersbasedon applyingBayes’theorem withstrong(naive)independenceassumptionsbetween thefeaturesinmachinelearning.Basicallywe canuse above theories andequationsforclassificationproblem.

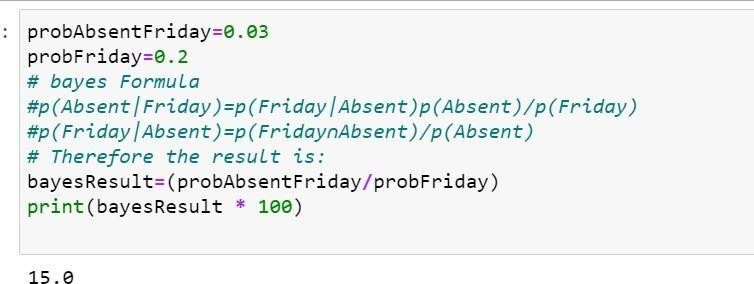
## SOURCECODE:

probAbsentFriday=0.03 probFriday=0.2

# bayes Formula#p(Absent|Friday)=p(Friday|Absent)p(Absent)/p(Friday)

#p(Friday|Absent)=p(Friday∩Absent)/p(Absent)#Thereforetheresultis:bayesResult=(probAbsentFriday/probFriday)print(bayesResult\*100)

**Output**:15



## 1.5.VIVAQUESTIONS&ANSWERS

1. **WhatareBayesianNetworks(BN)?**

Bayesian Network is used to represent the graphical model for probability relationship among a set ofvariables. Bayes’ theorem is a way to figure out [conditional probability.](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/conditional-probability-definition-examples/) Conditional probability is theprobability of an event happening, given that it has some relationship to one or more other events. Forexample, your probability of getting a parkingspace is connected to the time of day you park, whereyou park, and what conventions are going on at any time. Bayes’ theorem is slightly more nuanced. In anutshell,itgivesyouthe actual[probability](https://www.statisticshowto.com/probability-and-statistics/probability-main-index/)ofan**event**giveninformationabout**tests**.

* + “Events” Are different from “tests.” For example, there is a **test** for liver disease, but that’sseparate fromthe**event**ofactuallyhavingliverdisease.
  + **Tests are flawed**: just because you have a positive test does not mean you actually have thedisease. Many tests have a high[false positive rate](https://www.statisticshowto.com/false-positive-definition-and-examples/). **Rare events tend to have higher false positiverates** than more common events. We’re not just talking about medical tests here. For example, spamfiltering can have high false positive rates. Bayes’ theorem takes the test results and calculates your *realprobability*thatthetesthasidentifiedthe event.

### Can you giveanyrealtimeexampleusingBayes’Theorem (liverdisease).

You might be interested in finding out a patient’s probability of having liver disease if they are analcoholic.“Being analcoholic”isthe**test**(kindoflike alitmustest)forliverdisease.

* + **A**couldmeantheevent“Patienthasliverdisease.”Pastdatatellsyouthat10%ofpatientsenteringyourclinichaveliverdisease.P(A)= 0.10.
  + **B**couldmeanthelitmustestthat“Patientisanalcoholic.”Fivepercentoftheclinic’spatientsarealcoholics.P(B)=0.05.
  + You mightalsoknow that among those patients diagnosed with liver disease,7% are alcoholics.Thisisyour**B|A:**theprobabilitythatapatientisalcoholic,giventhattheyhaveliverdisease,is7%.

Bayes’theoremtellsyou:

## P(A|B)=(0.07\*0.1)/0.05=0.14

Inotherwords, ifthepatientisanalcoholic,theirchancesofhavingliverdiseaseis0.14(14%).

This is a large increase from the 10% suggested by past data. But it’s still unlikely that any particularpatienthasliverdisease.

1. **Bayes’TheoremExamples2:*whatistheprobabilitythattheywillbeprescribedpainpills?***

Another way to look at the theorem is to say that one event follows another. Above I said “tests” and“events”, but it’s also legitimate to think of it as the “first event” that leads to the “second event.”There’snoonerightwaytodothis:usetheterminologythatmakesmostsensetoyou.

In a particular pain clinic, 10% of patients are prescribed narcotic pain killers. Overall,five percent ofthe clinic’s patients are addicted to narcotics (including pain killers andillegal substances).Out of allthe people prescribed pain pills,8% are addicts.*If a patient is an addict,what is the probability thattheywillbeprescribedpainpills?*

**Step 1:Figure out what your event “A” is from the question.** That information is in the italicized partof this particular question. The event that happens first (A) is being prescribed pain pills. That’s given as10%.

**Step 2:Figure out what your event “B” is from the question.** That information is also in the italicizedpartofthisparticularquestion.EventBisbeinganaddict. That’s givenas 5%.

**Step 3:Figure out what the probability of event B (Step 2) given event A (Step 1)**. In other words,find what (B|A) is. We want to know “Given that people are prescribed pain pills, what’s the probabilitytheyareanaddict?”Thatis giveninthequestionas 8%,or.8.

## Step4:Insert your answers from Steps 1, 2 and 3 into the formula and solve.P(A|B)= P(B|A)\*P(A)/P(B)=(0.08\*0.1)/0.05= 0.16

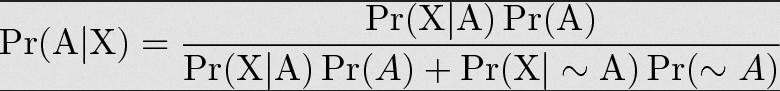
Theprobabilityofanaddictbeingprescribedpainpillsis0.16(16%).

### Bayes’ TheoremExamples3:theMedicalTestifaperson getsa positivetest result.

*what arethe oddsthey actually havethegeneticdefect?*

Aslightlymorecomplicatedexampleinvolvesamedicaltest(inthiscase,agenetictest):

There are **several forms of Bayes’ Theorem** out there, and they are all equivalent (they are just writtenin slightly different ways). In this next equation, “X” is used in place of “B.” In addition,you’ll seesome changes in the denominator. The proof of why we can rearrange the equation like thisisbeyondthe scope of this article (otherwise it would be 5,000 words instead of 2,000!). However, if you comeacross a question involving medical tests, you’ll likely be using this alternative formula to find theanswer:



1%ofpeople have a certain[genetic defect](https://www.genome.gov/10001204).

90%oftestsforthegenedetectthedefect(truepositives).9.6%ofthetestsare[falsepositives.](https://www.statisticshowto.com/false-positive-definition-and-examples/)

Ifapersongetsapositivetestresult,**whataretheodds theyactuallyhavethegeneticdefect?**

ThefirststepintosolvingBayes’theoremproblemsistoassignletterstoevents:

* + A=chanceofhavingthefaultygene.Thatwasgiveninthequestionas1%.That alsomeanstheprobabilityof*not*havingthegene(~A)is 99%.
  + X=Apositivetestresult.

So:

1. P(A|X)=Probabilityofhavingthegenegivenapositivetestresult.
2. P(X|A)=Chanceofapositivetestresultgiventhatthepersonactuallyhasthegene.Thatwasgiveninthequestionas 90%.
3. p(X|~A)=Chanceofapositivetestiftheperson*doesn’t*havethegene.Thatwasgiveninthequestionas 9.6%

Nowwehavealloftheinformationweneedtoputintotheequation:

P(A|X)=(.9\*.01)/(.9\*.01+.096\*.99)=0.0865(8.65%).

Theprobabilityofhavingthefaultygeneonthetestis 8.65%.

## 5.Giventhefollowingstatistics,whatistheprobability thatawomanhascancerifshehasapositive mammogramresult?

* + Onepercentofwomenover50havebreastcancer.
  + Ninetypercentofwomenwhohavebreastcancer testpositiveonmammograms.
  + Eightpercentofwomenwillhavefalsepositives.

Step 1: Assign events to A or X. You want to know what a woman’s probability of having cancer is,given a positive mammogram. For this problem, actually having cancer is A and a positive test result isX.

Step 2: List out the parts of the equation (this makes it easier to work the actual equation):P(A)=0.01

P(~A)=0.99P(X|A)=0.9P(X|~A)=0.08

Step 3: Insert the parts into the equation and solve. Note that as this is a medical test, we’re using theformoftheequationfromexample#2:

(0.9\*0.01)/((0.9\*0.01)+(0.08\*0.99)=0.10.

Theprobabilityofawomanhavingcancer,givenapositivetestresult,is10%.

# PROGRAM2

## EXTRACTTHEDATAFROMDATABASEUSINGPYTHON

You’ll learn thefollowingMySQL SELECT operationsfrom Python using a‘MySQLConnectorPython’module.

* + ExecutetheSELECTqueryandprocesstheresultsetreturnedbythequeryinPython.
  + UsePython**variablesinawhereclause**ofaSELECTquerytopassdynamicvalues.
  + Use[fetchall(),fetchmany(),andfetchone()](https://pynative.com/python-cursor-fetchall-fetchmany-fetchone-to-read-rows-from-table/)methods of acursorclasstofetch all orlimitedrows fromatable.
  + PythonSelectfromMySQLTable

Thisarticledemonstrateshowto**selectrowsofa MySQLtablein Python**.

You’lllearnthefollowingMySQLSELECToperationsfromPythonusinga‘MySQLConnectorPython’module

.

* + ExecutetheSELECTqueryandprocesstheresultsetreturnedbythequeryinPython.
  + UsePython**variablesinawhereclause**ofaSELECTquerytopassdynamicvalues.
  + Use[fetchall(),fetchmany(),andfetchone()](https://pynative.com/python-cursor-fetchall-fetchmany-fetchone-to-read-rows-from-table/)methodsofacursorclasstofetchallorlimitedrowsfromatable.
  + StepstofetchrowsfromaMySQLdatabasetableFollowthesesteps:–

## SOURCECODE:

importpymysql

defmysqlconnect():

# ToconnectMySQLdatabaseconn=pymysql.connect(

host='localhost', user='root',password = "pass",db='College',

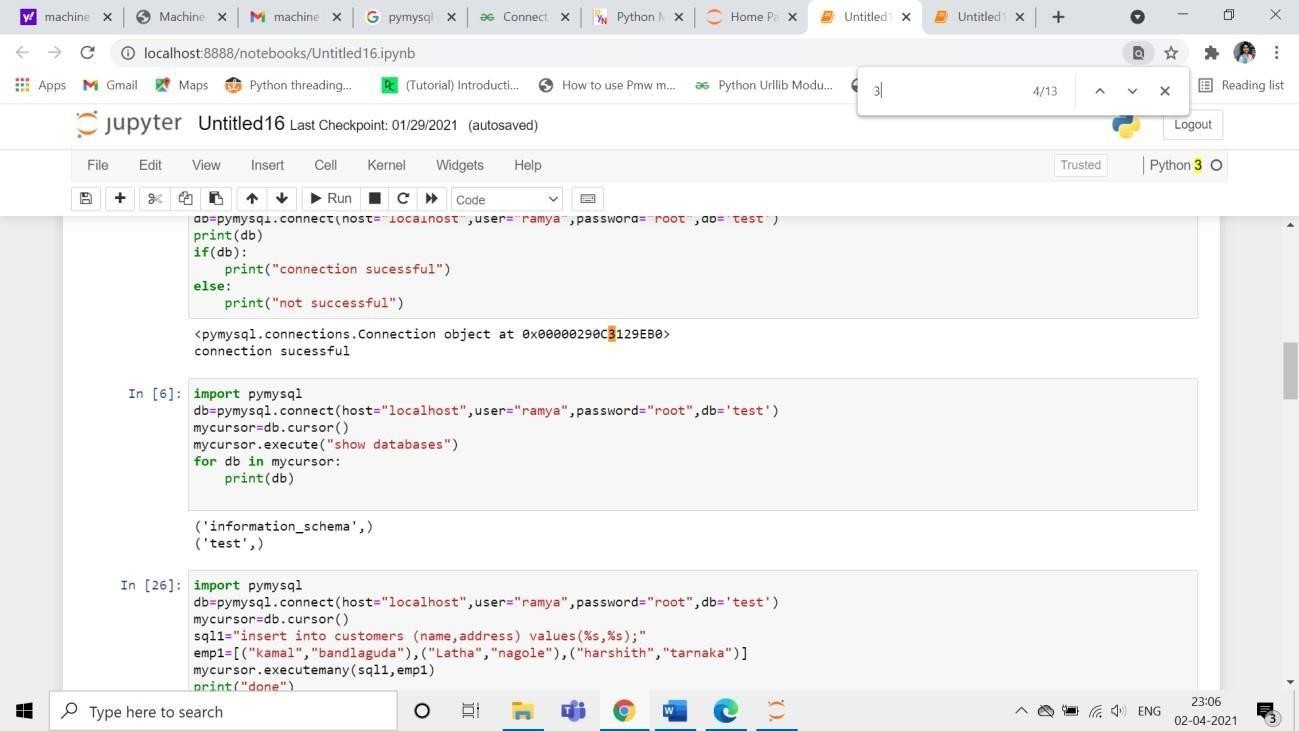
)

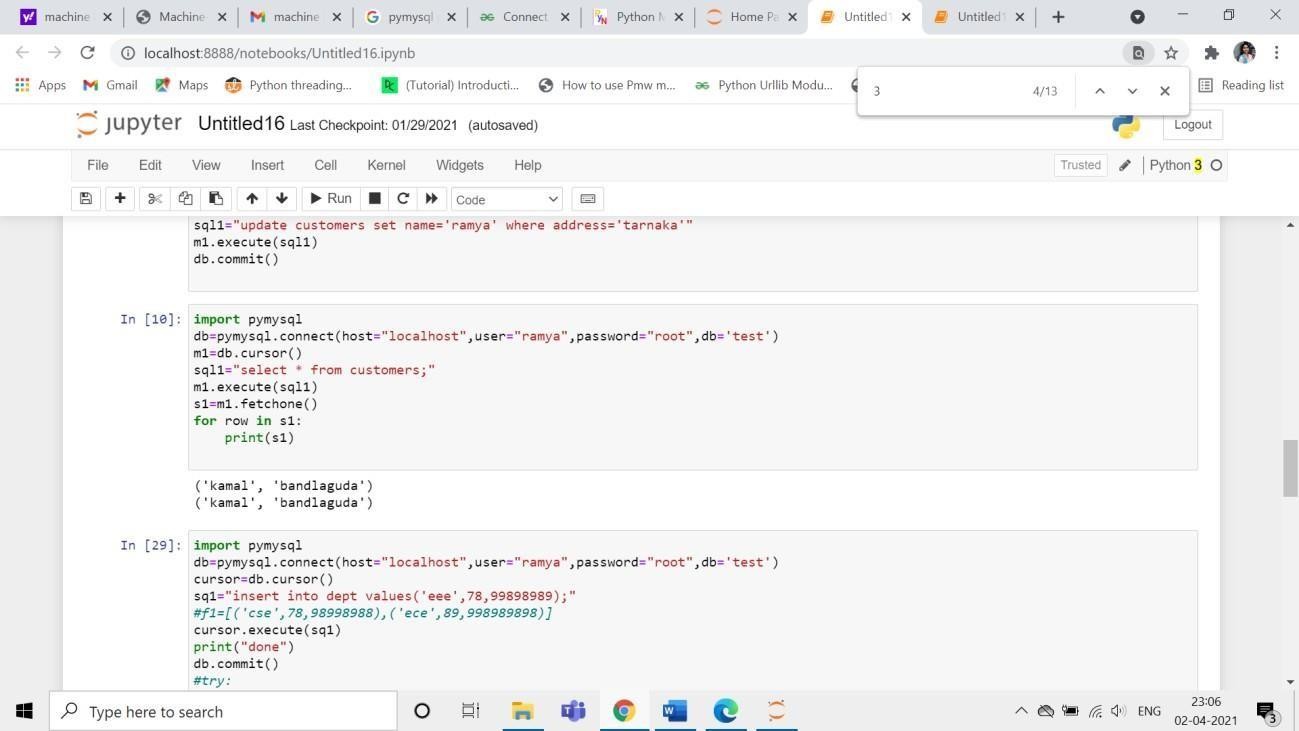
cur=conn.cursor()cur.execute("select@@version")output=cur.fetchall()print(output)

#Toclosetheconnection conn.close(

)

**OUTPUT:**





## VIVAQUESTIONS ANDANSWERS

1. **Howtoselect fromaMySQLtableusing Python?**

## ConnecttoMySQLfromPython

Refer to [Python MySQL database connection](https://pynative.com/python-mysql-database-connection/)to connect to MySQL database from Python usingMySQL Connectormodule

**Define**aSQLSELECTQuery

Next, prepare a SQL SELECT query to fetch rows from a table. You can select all or limited rows basedon yourrequirement.If the where conditionisused,thenitdecides the numberof rows tofetch.For example, SELECT col1, col2,…colnN FROM MySQL\_table WHEREid = 10;. This will return row number10.

**Get**CursorObjectfromConnection

Next, useaconnection.cursor()methodtocreateacursorobject.Thismethod createsanewMySQLCursor

object.

**Execute** the SELECT query using execute() methodExecutetheselectqueryusingthecursor.execute()method.**Extract**allrows froma result

AftersuccessfullyexecutingaSelectoperation,Usethe[fetchall()](https://pynative.com/python-cursor-fetchall-fetchmany-fetchone-to-read-rows-from-table/)methodofacursorobjecttogetallrowsfromaqueryresult.itreturns alistofrows.

**Iterate**eachrow

Iterate arowlistusingafor loopand accesseachrowindividually(Accesseachrow’scolumndatausingacolumnname orindexnumber.)

## Closethecursorobjectanddatabaseconnectionobject

usecursor.clsoe()andconnection.clsoe()methodtocloseopenconnectionsafteryourworkcompletes.

## Whatarethemethodstofetchdatareturnedbyacursor.execute()

* + cursor.fetchall()tofetchallrows
  + cursor.fetchone()tofetchasingle row
  + cursor.fetchmany(SIZE)tofetchlimitedrows

# PROGRAM 3

1. **IMPLEMENTK-NEARESTNEIGHBORSCLASSIFICATIONUSINGPYTHONAIM:**ToImplementk-neighboursclassificationusingpython

## DESCRIPTION:

Thisalgorithmisusedtosolvetheclassificationmodelproblems.K-nearestneighbororK-NNalgorithm basically creates an imaginary boundary to classify the data.Whennew data points comein,thealgorithmwilltrytopredictthattothenearestoftheboundaryline.

Therefore,largerkvaluemeanssmothercurvesofseparationresultinginlesscomplexmodels.Whereas,smallerkvalue tendstooverfitthedata andresultingincomplexmodels.

It’s very important to have the right k-value when analyzing the dataset to avoid over fitting and underfittingofthedataset.

## KNNMODELREPRESENTATION:

The model representation for KNN is the entire training dataset.It is as simple as that.KNNhasnomodelotherthanstoringtheentiredataset,sothereisnolearningrequired.

Efficientimplementationscanstorethedatausingcomplexdatastructureslike[k-dtrees](https://en.wikipedia.org/wiki/K-d_tree)tomakelook-upandmatchingofnew patterns duringpredictionefficient.

Because the entire training dataset is stored, you may want to think carefully about the consistency of yourtraining data. It might be a good idea to curate it, update itoften as new data becomes available andremove erroneous andoutlierdata.

* The**k**-**nearestneighboralgorithm**isimportedfromthescikit-learnpackage.
* Createfeatureandtargetvariables.
* Splitdataintotrainingand testdata.
* Generate a**k**-**NN**modelusing**neighbors**value.
* Trainorfitthedataintothemodel.
* Predictthefuture.

## SOURCECODE:

#Importnecessarymodules

fromsklearn.neighborsimportKNeighborsClassifierfromsklearn.model\_selectionimporttrain\_test\_splitfromsklearn.datasetsimportload\_iris

# Loading data irisData =load\_iris()

#CreatefeatureandtargetarraysX=irisData.data

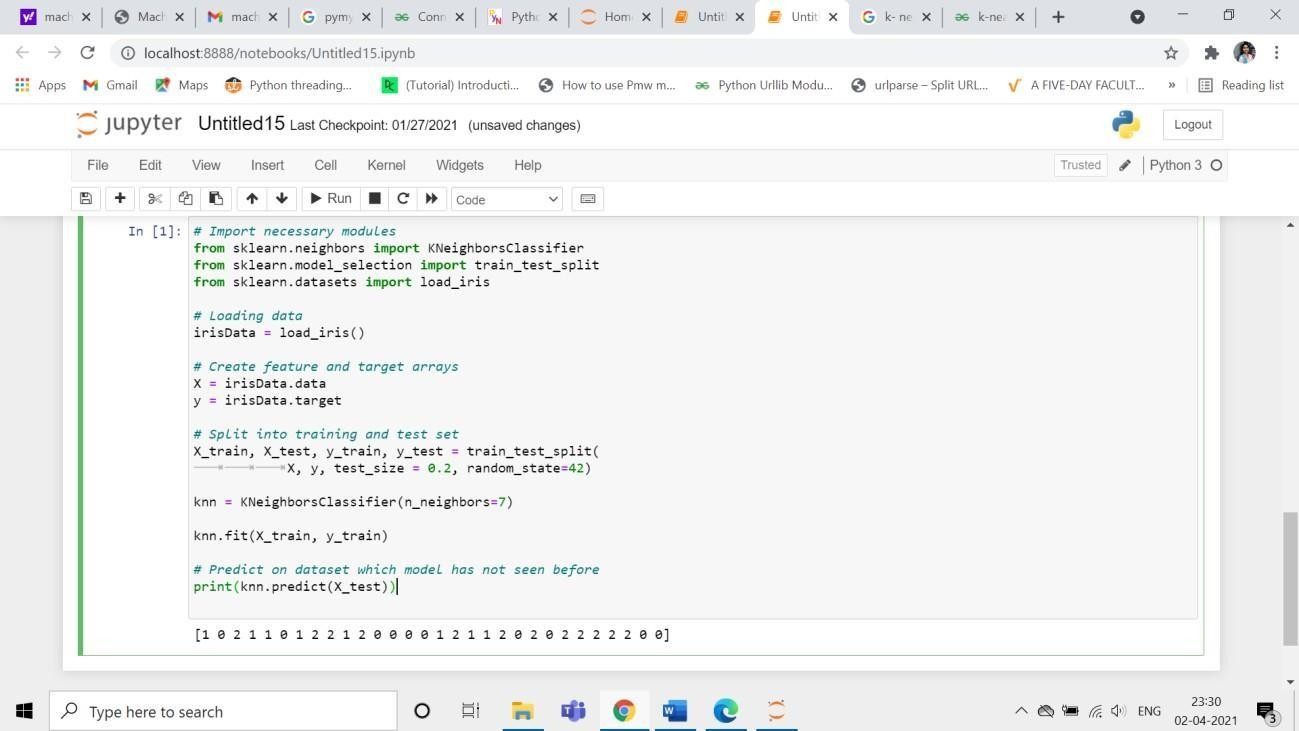
y=irisData.target

#Splitintotrainingandtestset

X\_train,X\_test,y\_train,y\_test=train\_test\_split(X, y, test\_size = 0.2, random\_state=42) knn =KNeighborsClassifier(n\_neighbors=7)knn.fit(X\_train,y\_train)

#Predictondataset whichmodelhasnot seenbeforeprint(knn.predict(X\_test))

## 3.5OUTPUT:



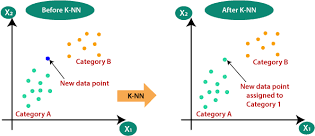
**VAQUESTIONS&ANSWERS**

## WhatistheKNNAlgorithm?

**KNN(K-nearestneighbors)**isa**supervised**learningand**non-parametric**algorithmthatcanbeusedtosolve bothclassificationandregressionproblemstatements.

It uses data in which there is a target column present **i.e, labeled data** to model a function to produce anoutputfortheunseendata.It usestheEuclideandistanceformulatocomputethedistancebetweenthedatapoints forclassificationorprediction.

Themainobjectiveofthisalgorithmisthatsimilardatapointsmustbeclosetoeachothersoitusesthedistance tocalculatethe similarpointsthatareclosetoeachother.



## WhyisKNNanon-parametricAlgorithm?

Theterm“**non-parametric”**referstonotmakinganyassumptionsontheunderlyingdatadistribution.Thesemethods donothaveanyfixednumbers ofparametersinthemodel.

Similarly in KNN, the model parameters grow with the training data by considering each training case as aparameterofthemodel.So,KNNisa non-parametricalgorithm.

## Whatis“K”intheKNNAlgorithm?

Krepresentsthenumberofnearestneighborsyouwanttoselecttopredicttheclassofagivenitem,whichiscomingas anunseendatasetfor themodel.

## Whyistheoddvalue of“K”preferred overeven valuesintheKNNAlgorithm?

TheoddvalueofKshouldbepreferredoverevenvaluesinordertoensurethattherearenotiesinthevoting.Ifthesquarerootofanumberofdatapointsiseven,thenaddorsubtract1toittomakeitodd.

## Howdoes theKNNalgorithm make the predictions onthe unseendataset?

Thefollowingoperationshavehappenedduringeachiterationofthealgorithm.Foreachoftheunseenortestdatapoint,the kNNclassifiermust:

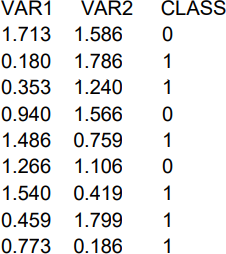
**Step-1:**Calculatethedistancesoftestpointtoallpointsinthetrainingsetandstorethem

**Step-2:** Sort the calculated distances in increasing order**Step-3:**StoretheKnearestpointsfromourtrainingdataset**Step-4:**Calculatetheproportionsofeachclass

**Step-5:**Assigntheclasswiththehighestproportion

# PROGRAM-4

## 4.Given the following data, which specify classifications for nine combinations of VAR1andVAR2predict aclassificationforacasewhereVAR1=0.906 andVAR2=0.606,usingthe resultofk-means clusteringwith3means (i.e.,3centroids)



**SOURCECODE:**

fromsklearn.clusterimportKMeansimportnumpyasnp

X=np.array([[1.713,1.586],[0.180,1.786],[0.353,1.240],

[0.940,1.566],[1.486,0.759],[1.266,1.106],[1.540,0.419],[0.459,1.799],[0.773,0.186]])y=np.array([0,1,1,0,1,0,1,1,1])

kmeans=KMeans(n\_clusters=3,random\_state=0).fit(X,y)

## 4.2.OUTPUT:

kmeans.predict([[0.906,0.606]])

## 4.3VIVAQUESTIONS&ANSWERS

1. **WhatisK meansClustering Algorithm?**

K Means algorithm is a centroid-based clustering (unsupervised) technique. This technique groups thedataset into k differentclusters having an almost equal number of points. Each of the clusters has acentroidpointwhichrepresentsthemeanofthedatapointslyinginthatcluster.

The idea of the K-Means algorithm is to find k-centroid points and every point in the dataset will belong toeitherofthek-sets havingminimumEuclideandistance.

## IsFeatureScaling requiredfortheKmeansAlgorithm?

**Yes,** K-Means typically needs to have some form of normalization done on the datasets to work properlysinceitis sensitive toboththemeanandvarianceofthe datasets.

For performing feature scaling, generally, **StandardScaler**is recommended, but depending on the specificuse cases,othertechniquesmightbemore suitableaswell.

**For Example,**let’s have 2 variables, named ageand salary where ageis in the range of 20 to 60 andsalary is in the range of 100-150K, since scales of these variables are different so when these variables aresubstituted in the euclidean distance formula, then the variable which is on the large scale suppresses thevariable which is on the smaller scale. So, the impact of age will not be captured very clearly. Hence, youhavetoscale thevariables tothe samerange using**StandardScaler,Min-MaxScaler**,etc.

## Whichmetricscan youuseto findtheaccuracyoftheKmeansAlgorithm?

There does not exist a correct answer to this question as k means being an unsupervised learning techniquedoes not discuss anything about the output column. As a result, one can not get the accuracy number orvaluesfromthealgorithmdirectly.

## WhataretheadvantagesanddisadvantagesoftheKmeansAlgorithm?Advantages:

* Easytounderstandandimplement.
* Computationallyefficientforbothtrainingandprediction.
* Guaranteedconvergence.

**Disadvantages:**

* Weneedtoprovidethenumberofclustersasaninputvariabletothealgorithm.
* Itisverysensitivetotheinitializationprocess.
* Good at clustering when we are dealing with spherical cluster shapes, but it will perform poorlywhendealingwithmorecomplicatedshapes.
* DuetotheleveragingoftheEuclideandistancefunction,itissensitive tooutliers.

## Whatarethewaystoavoidtheproblemofinitialization sensitivity intheKmeansAlgorithm?

Therearetwowaystoavoidtheproblemofinitializationsensitivity:

* **RepeatKmeans:** Itbasically repeatsthealgorithm againandagain alongwithinitializingthecentroidsfollowedbypickinguptheclusterwhichresultsinthesmallintraclusterdistanceandlarge

interclusterdistance.

* **KMeans++:**Itisasmartcentroidinitializationtechnique.

Amongsttheabovetwotechniques,K-Means++isthebestapproach.

# PROGRAM 5

## TheFollowing Training ExamplesMap DescriptionsOfIndividualsOntoHigh,MediumAndLowCredit-Worthiness.

medium skiing design single twenties no ->highRiskhighgolftradingmarriedfortiesyes->lowRisk

lowspeedwaytransportmarried thirtiesyes->medRiskmedium football banking single thirties yes ->lowRiskhighflyingmediamarriedfiftiesyes->highRisk

low football security single twenties no ->medRiskmedium golf media single thirties yes ->medRiskmedium golf transport married forties yes ->lowRiskhigh skiing banking single thirties yes ->highRisk lowgolfunemployedmarriedfortiesyes->highRisk

## SOURCECODE:

Inputattributes are (from lefttoright)income,recreation,job,status,age-group, home-owner.Findtheunconditionalprobabilityof`golf'andtheconditionalprobabilityofsingle' given

`medRisk' in thedataset?totalRecords=10numberGolfRecreation=4

probGolf=numberGolfRecreation/totalRecordsprint("Unconditional probability of golf: ={}".format(probGolf)) #conditionalprobabilityof`single'given`medRisk'

#bayesFormula#p(single|medRisk)=p(medRisk|single)p(single)/p(medRisk)#p(medRisk|single)=p(medRisk ∩single)/p(single)

#Thereforetheresultis:

numberMedRiskSingle=2numberMedRisk=3

probMedRiskSingle=numberMedRiskSingle/totalRecordsprobMedRisk=numberMedRisk/totalRecordsconditionalProbability=(probMedRiskSingle/probMedRisk)

print("ConditionalprobabilityofsinglegivenmedRisk:={}".format(conditionalProbability))

## OUTPUT:

Unconditionalprobabilityofgolf:=0.4

ConditionalprobabilityofsinglegivenmedRisk:=0.6666666666666667

## VIVAQUESTIONS&ANSWERS:

* 1. **HowKmeans++clusteringAlgorithmworks?**

K Means++ algorithm is a smart technique for centroid initialization that initialized one centroid whileensuringtheotherstobefarawayfromthechosenoneresultinginfasterconvergence.

Thestepstofollowforcentroid initializationare:

**Step-1:**Pickthefirstcentroidpointrandomly.

**Step-2:**Computethedistanceofallpointsinthe datasetfromtheselectedcentroid.Thedistanceofxipointfromthefarthestcentroidcanbecalculatedbythegivenformula:

distance calculation

## where,

**di:**Distanceofxipointfromthefarthestcentroid

**m:**numberofcentroidsalreadypicked

**Step-3:**Makethepointxiasthenewcentroidthatishavingmaximumprobabilityproportionaltodi

**Step-4:**Repeattheabovelasttwostepstillyoufindkcentroids.

## Whatisthetraining andtestingcomplexityoftheKmeansAlgorithm?

**TrainingcomplexityintermsofBig-Onotation:**

IfweuseLloyd’salgorithm,thecomplexityfortrainingis:**“K\*I\*N\*M”**

where,

**K:** It represents the number of clusters**I:**Itrepresentsthenumberofiterations**N:**Itrepresentsthe sample size

**M:**Itrepresentsthenumberofvariables

**Conclusion:** ThereisasignificantImpactoncappingthenumberofiterations.

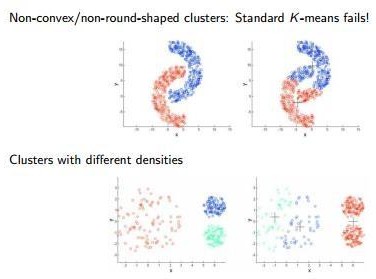
## PredictingcomplexityintermsofBig-Onotation:

**“K\*N\*M”**

Predictionneedstobecomputedforeachrecord,thedistancetoeachclusterandassignedtothenearestones.

## Isitpossiblethat theassignmentofdatapointstoclustersdoesnotchangebetweensuccessiveiterations intheKmeans Algorithm?

WhentheK-Meansalgorithmhasreachedthelocalorglobalminima,itwillnotchangetheassignmentofdatapointstoclustersfor twosuccessiveiterationsduringthealgorithmrun.



* 1. **HowtoChooseKValueinK-Means:**

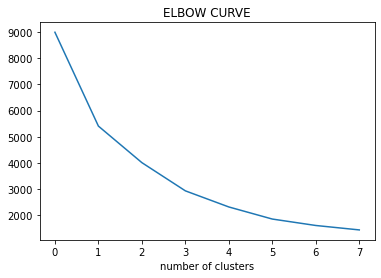
## 1.Elbowmethod

steps:

step1: compute clustering algorithm for different values of k.forexamplek=[1,2,3,4,5,6,7,8,9,10]

step2: for each k calculate the within-cluster sum of squares(WCSS).step3: plotcurveofWCSSaccordingtothe numberofclusters.

step4: Thelocation ofbendin theplotisgenerallyconsideredanindicatoroftheapproximatenumberofclusters.



**Page30**

## WhatarethepracticalConsiderationsInK-Means:

* AchoosingnumberofClustersinAdvance(K).
* StandardizationofData(scaling).
* CategoricalData(canbesolvedwithK-Mode).
* ImpactofinitialCentroidsandOutliers.

## WhydoyoupreferEuclideandistance overManhattandistanceintheKmeansAlgorithm?

Euclidean distanceis preferred over Manhattan distance since Manhattan distance calculates distanceonlyverticallyorhorizontallyduetowhichithasdimensionrestrictions.

On the contrary, Euclidean distance can be used in any space to calculate the distances between the datapoints. Since in K means algorithm the data points can be present in any dimension, so Euclideandistanceis amore suitableoption.

**Page31**

# PROGRAM 6:

## 6.IMPLEMENTLINEARREGRESSIONUSINGPYTHON.AIM:

ToImplementlinearregressionusingpython.

## DESCRIPTION:

**Regression:**Regressionanalysisisoneofthemostimportantfieldsinstatisticsandmachinelearning.Therearemanyregressionmethodsavailable.Linearregressionisoneofthem

## WhatIsRegression?

Regressionanalysisisoneofthe mostimportantfieldsinstatisticsand machinelearning.Thereare many regression methods available. Linear regression is one ofthem.Regression searchesforrelationships amongvariables.Forexample,youcan observeseveralemployeesofsomecompanyandtrytounderstandhowtheirsalariesdependonthe**features**,suchasexperience,levelofeducation,role,citytheyworkin,andsoon.Thisisaregressionproblemwheredata relatedtoeachemployee representone**observation**.Thepresumptionisthattheexperience,education,role,andcityaretheindependentfeatures,whilethesalarydependsonthem.Generally,inregressionanalysis,youusuallyconsidersomephenomenonofinterestandTherehaveanumberofobservations.Eachobservationhastwoormore features. Following the assumptionthat(atleast)oneofthefeaturesdependsontheothers,youtrytoestablisharelationamongthem.youneedtofindafunctionthatmapssomefeaturesorvariablestootherssufficientlywell.Thedependentfeatures are calledthe**dependentvariables**,**outputs**,or**responses**.Theindependentfeatures arecalledthe **independentvariables**,**inputs**,or**predictors**.

## LinearRegression:

Linearregressionisprobablyoneofthemostimportantandwidelyusedregressiontechniques. It’s among the simplest regression methods. One of its main advantages is theeaseofinterpretingresults.Whenimplementinglinearregressionofsomedependentvariable

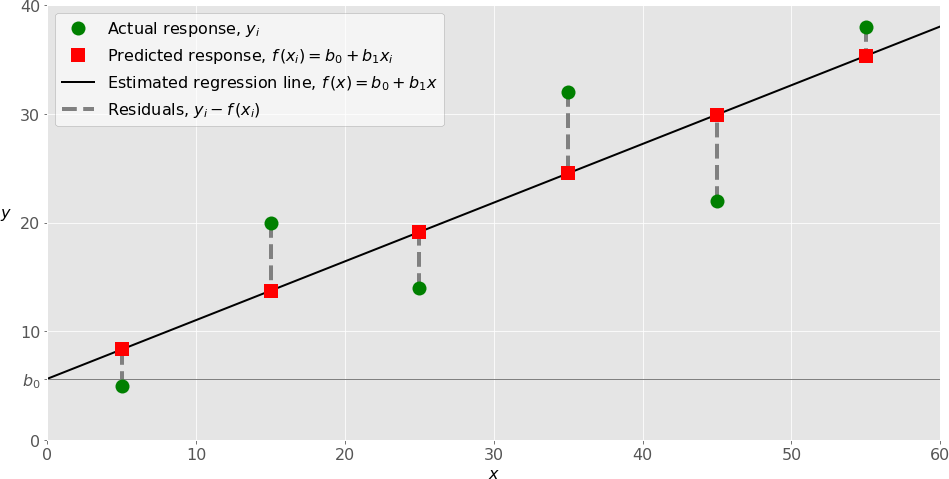
𝑦on theset of independent variablesx= (x1…xrᵣ),whereristhe number ofpredictors,you assume a linear relationship between y andx: y= β0 + β1x1 + ……+ βrxr+ ϵ. This equation isthe**regressionequation**.β0,,βrᵣarethe**regressioncoefficients**,and𝜀isthe**randomerror**.

Linear regression calculates the **estimators** of the regression coefficients or simply the **predictedweights**,denotedwithb1…br.Theydefinethe**estimatedregressionfunctionf**(x)=b0+b1x1+

….+brxrThisfunctionshouldcapturethedependenciesbetweentheinputsandoutputsufficientlywell.

## SimpleLinearRegression

Thefollowingfigureillustratessimplelinearregression:



When implementing simple linear regression, you typically start with a given set of input-output(x-y)pairs(greencircles).Thesepairsareyour observations.Forexample,theleftmostobservation(greencircle)hastheinputx=5andtheactualoutput(response)y=5.Thenextonehasx=15andy= 20,andsoon.

Theestimatedregressionfunction(blackline)hastheequationf(x)=b0+b1x.Yourgoalistocalculatetheoptimalvaluesofthepredictedweightsb0andb1thatminimizeSSRanddeterminetheestimatedregressionfunction.Thevalueofb0,alsocalledthe**intercept**,showsthepointwheretheestimatedregressionlinecrossesthe y axis.Itisthevalueoftheestimatedresponse f(x)for x=0. Thevalue of b1determinesthe**slope**oftheestimatedregressionline.

The predicted responses (red squares) are the points on the regression line that correspond to theinput values. For example, for the input x= 5, the predicted response is f (5) = 8.33 (representedwiththeleftmostredsquare).

Theresiduals(verticaldashedgraylines)canbecalculatedasy1ᵢ-f(x1ᵢ)=y1–b0–b1x1fori=1,

…,n . They are the distances between the green circles and red squares. When you implementlinear regression, you are actually trying to minimize these distances and make the red squares asclose tothepredefinedgreencircles aspossible.

## ImplementingLinearRegressioninPython

It’s time to start implementing linear regression in Python. Basically, all you should do is applythe properpackagesandtheirfunctions andclasses.

## PythonPackagesforLinearRegression

The package **NumPy**is a fundamental Python scientific package that allows many high-performanceoperationsonsingle-andmulti-dimensionalarrays.Italsooffersmanymathematical routines.Ofcourse,it’s opensource.

Thepackage**scikit-learn**isawidelyusedPythonlibraryformachinelearning,builtontopofNumPyandsomeotherpackages.Itprovidesthemeansforpreprocessingdata,reducingdimensionality,implementingregression,classification,clustering,andmore.LikeNumPy,scikit-learnis alsoopensource.

If you want to implement linear regression and need the functionality beyond the scope of scikit-learn, you should consider **statsmodels**. It’s a powerful Python package for the estimation ofstatisticalmodels,performingtests,andmore.It’sopensource as well.

## SimpleLinearRegressionWith scikit-learn

Let’s start with the simplest case, which is simple linear regression. Therearefivebasicstepswhenyou’reimplementinglinear regression:

1. Importthepackagesandclasses youneed.
2. Providedatatoworkwithandeventuallydoappropriatetransformations.
3. Createaregressionmodelandfititwithexistingdata.
4. Checktheresultsofmodelfittingtoknowwhetherthemodelissatisfactory.
5. Applythemodelforpredictions.

## SOURCECODE:

importnumpy as npimportmatplotlib.pyplotasplt

defestimate\_coef(x,y):

#numberofobservations/pointsn=

np.size(x)

#meanofxandyvector

m\_x,m\_y=np.mean(x),np.mean(y)

#calculatingcross-deviationanddeviationaboutxSS\_xy=np.sum(y\*x)-n\*m\_y\*m\_x

SS\_xx=np.sum(x\*x) -n\*m\_x\*m\_x

#calculatingregressioncoefficientsb\_1=

SS\_xy/SS\_xx

b\_0 = m\_y - b\_1\*m\_xreturn(b\_0,b\_1)

defplot\_regression\_line(x,y,b):

# plotting the actual points as scatter plotplt.scatter(x,y,color="m",

marker="o",s =30)

#predictedresponsevectory\_pred=b[0]+b[1]\*x

# plotting the regression lineplt.plot(x, y\_pred,color="g")

# putting labelsplt.xlabel('x')

plt.ylabel('y')

#functionto showplotplt.show()

defmain():

#observations

x=np.array([0,1, 2, 3,4, 5,6, 7, 8, 9])

y=np.array([1,3, 2, 5, 7, 8,8, 9, 10, 12])

#estimatingcoefficientsb=estimate\_coef(x,y)

print("Estimatedcoefficients:\nb\_0={}\

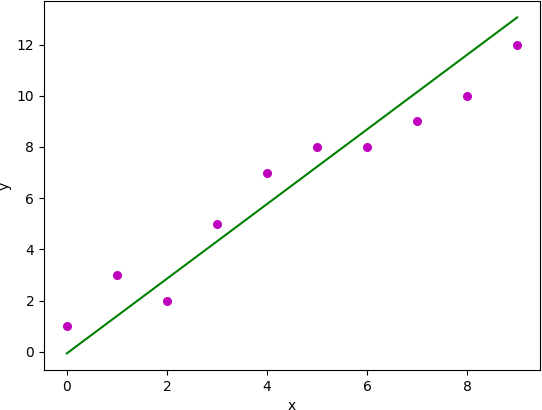
\nb\_1={}".format(b[0],b[1]))

# plotting regression lineplot\_regression\_line(x,y,b)

ifname == "main":main()

## OUTPUT:

Estimated coefficients:b\_0=-0.05862068965



## VIVAQUESTIONS ANDANSWERS

1. **WhatdoyoumeanbytheLogisticRegression?**

It’s a classification algorithm that is used where the target variable is of categorical nature. The mainobjective behind Logistic Regression is to determine the relationship between features and the probabilityofaparticularoutcome.

**For Example,** when we need to predict whether a student passes or fails in an exam given the number ofhoursspentstudying asafeature,thetargetvariable comprisestwovaluesi.e.passandfail.

Therefore,wecansolveclassificationproblemstatementswhichisasupervisedmachinelearningtechnique usingLogisticRegression.

## WhatarethedifferenttypesofLogisticRegression?

ThreedifferenttypesofLogistic Regressionareasfollows:

1. **BinaryLogisticRegression:**Inthis,thetargetvariablehasonlytwo2possibleoutcomes.

**ForExample,**0and 1,orpassandfailortrueandfalse.

1. **Multinomial Logistic Regression:** In this, the target variable can have three or more possible valueswithoutanyorder.

**ForExample,**Predictingpreferenceoffoodi.e.Veg,Non-Veg,Vegan.

1. **OrdinalLogisticRegression:**Inthis,thetargetvariablecanhavethreeormorevalueswithordering.

**ForExample,**Movieratingfrom1to5.

## Howdowehandlecategoricalvariablesin LogisticRegression?

Theinputs given toa Logistic Regression model need to be numeric.The algorithm cannothandlecategorical variables directly. So, we need to convert the categorical data into a numerical format that issuitableforthealgorithmtoprocess.

Each level of the categorical variable will be assigned a unique numeric value also known as a **dummyvariable.**These dummy variables are handledby the Logistic Regressionmodel in the samemannerasanyothernumericvalue.

## WhataretheassumptionsmadeinLogisticRegression?

SomeoftheassumptionsofLogisticRegressionareasfollows:

1. Itassumesthatthereis minimalor**nomulticollinearity**amongtheindependentvariablesi.e,predictors arenotcorrelated.
2. Thereshouldbealinearrelationshipbetweenthelogitoftheoutcomeandeachpredictor variable.Thelogitfunctionisdescribedas**logit(p)=log(p/(1-p))**,wherepistheprobabilityofthetargetoutcome.
3. Sometimesto predictproperly,itusuallyrequiresa**largesamplesize**.
4. TheLogisticRegressionwhichhas**binaryclassification**i.e,twoclassesassumethatthetargetvariableisbinary,andorderedLogistic Regressionrequires the targetvariable tobe ordered.

**Forexample,**Too Little,AboutRight,TooMuch.

1. Itassumesthereis**nodependency** betweentheobservations.

## *5.* CanwesolvethemulticlassclassificationproblemsusingLogisticRegression?IfYesthen How?

**Yes**, in order to deal with multiclass classification using Logistic Regression, the most famous method isknownastheone-vs-allapproach.Inthisapproach,anumberofmodelsaretrained,which isequaltothenumberofclasses.Thesemodels workina specificway.

**For Example,** the first model classifies the data point depending on whether it belongs to class 1 or someotherclass(notclass1);thesecondmodelclassifiesthedatapointintoclass2orsomeotherclass(notclass2) and so-onforallotherclasses.So,inthismanner,eachdatapointcanbechecked overalltheclasses.

# PROGRAM 7

## IMPLEMENTNAÏVEBAYESTHEOREM TOCLASSIFYTHEENGLISHTEXTAIM:

ToImplementnaïvebaye’stheoremtoclassifytheEnglishtext

## DESCRIPTION:

The challenge of text classification is to attach labels to bodies of text, e.g., tax document, medical form,etc. based on the text itself. For example, think of your spam folder in your email. How does your emailprovider know that a particular message is spam or “ham” (not spam)? We’ll take a look at one naturallanguage processingtechniquefor textclassificationcalledNaiveBayes

## SOURCECODE:

importpandasaspd

fromsklearn.model\_selection import train\_test\_splitfromsklearn.feature\_extraction.textimport CountVectorizerfromsklearn.naive\_bayesimportMultinomialNB

fromsklearn.metricsimportaccuracy\_score,confusion\_matrix,precision\_score,recall\_score

msg = pd.read\_csv('document.csv', names=['message', 'label'])print("Total Instances of Dataset: ", msg.shape[0])msg['labelnum'] =msg.label.map({'pos': 1,'neg':0})

X = msg.messagey=msg.labelnum

Xtrain,Xtest,ytrain,ytest=train\_test\_split(X,y)count\_v=CountVectorizer()

Xtrain\_dm = count\_v.fit\_transform(Xtrain)Xtest\_dm=count\_v.transform(Xtest)

df=pd.DataFrame(Xtrain\_dm.toarray(),columns=count\_v.get\_feature\_names())

clf = MultinomialNB()clf.fit(Xtrain\_dm,ytrain)pred=clf.predict(Xtest\_dm)

print('AccuracyMetrics:')

print('Accuracy:',accuracy\_score(ytest,pred))print('Recall:',recall\_score(ytest,pred))print('Precision:',precision\_score(ytest,pred))

print('ConfusionMatrix:\n',confusion\_matrix(ytest,pred))

## document.csv:

Ilovethissandwich,posThisisanamazingplace,pos

I feel very good about these beers,posThisismybestwork,pos

Whatanawesomeview,pos

Idonotlikethisrestaurant,negIamtiredofthis stuff,neg

I can't deal with this,negHeismyswornenemy,negMy

bossishorrible,neg

Thisisanawesome place,pos

I do not like the taste of this juice,negIlovetodance,pos

Iamsickandtiredofthisplace,negWhata greatholiday,pos

Thatisabadlocalitytostay,neg

We willhave goodfuntomorrow,posIwenttomyenemy'shousetoday,neg

## OUTPUT:

Total Instances of Dataset: 18AccuracyMetrics:

Accuracy:0.6

Recall: 0.6666666666666666

Precision:0.6666666666666666

Confusion Matrix:[[11]

[12]]

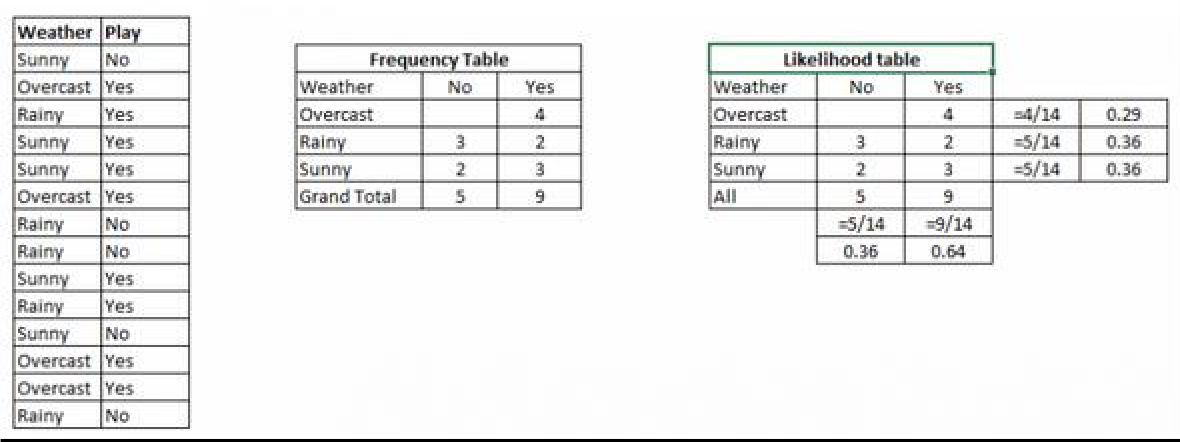
## VIVAQUESTIONS & ANSWERS

1. **HowNaiveBayesalgorithmworks?**

Let’s understand it using an example. Below I have a training data set of weather and corresponding targetvariable ‘Play’ (suggesting possibilities of playing). Now, we need to classify whether players will play ornotbasedonweathercondition.Let’sfollowthebelow stepstoperformit.

Step1:Convertthedatasetintoafrequencytable

Step2:CreateLikelihoodtablebyfindingtheprobabilitieslikeOvercastprobability=0.29andprobabilityofplayingis 0.64.



Step3:Now,use[NaiveBayesian](https://courses.analyticsvidhya.com/courses/naive-bayes?utm_source=blog&utm_medium=naive-bayes-explained)equationtocalculatetheposterior probabilityfor eachclass.Theclasswiththehighestposteriorprobabilityis the outcomeofprediction.

**Problem:** Players will play if weather is sunny. Is this statement is correct?Wecansolveitusingabovediscussedmethodofposteriorprobability.

P(Yes|Sunny)=P(Sunny|Yes) \* P(Yes) /P(Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P( Yes)= 9/14 = 0.64Now,P(Yes|Sunny)= 0.33\*0.64/0.36= 0.60,whichhashigherprobability.

NaiveBayesusesasimilarmethodtopredicttheprobabilityofdifferentclassbasedonvariousattributes.Thisalgorithmismostlyusedintextclassificationandwithproblemshavingmultipleclasses.

## ApplicationsofNaiveBayesAlgorithms

* **Real time Prediction:** Naive Bayes is an eager learning classifier and itis sure fast. Thus, it couldbe usedformakingpredictionsinrealtime.
* **Multi class Prediction:** This algorithm is also well known for multi class prediction feature. Herewe canpredictthe probabilityofmultipleclassesoftargetvariable.
* **Text classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classifiers mostly used intext classification (due to better result in multi class problems and independence rule) have higher successrate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail)andSentimentAnalysis(insocialmediaanalysis,toidentifypositiveandnegativecustomersentiments)
* **Recommendation System:** Naive Bayes Classifier and [Collaborative Filtering](https://en.wikipedia.org/wiki/Collaborative_filtering) together builds aRecommendationSystemthatusesmachinelearninganddataminingtechniquestofilterunseeninformationandpredictwhethera userwouldlike a givenresourceornot

# PROGRAM 8

## IMPLEMENTANALGORITHMTODEMONSTRATETHESIGNIFICANCEOFGENETICALGORITHM

**AIM:**

ToImplementanalgorithmtodemonstratethesignificanceofgeneticalgorithm

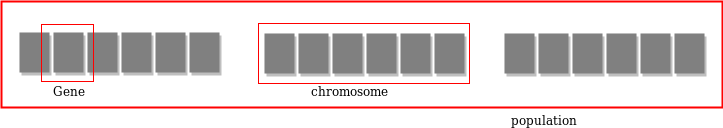
## DESCRIPTION:

Genetic Algorithms (GAs) are adaptive heuristic search algorithms that belong to thelarger part of evolutionary algorithms. Genetic algorithms are based on the ideas of naturalselectionandgenetics.Theseareintelligentexploitationof random searchprovidedwithhistorical data to direct the search into the region of better performance in solution space. Theyare commonly used to generate high-quality solutions for optimization problems and searchproblems.Geneticalgorithmssimulatetheprocessofnaturalselectionwhichmeansthosespecies who can adapt to changes in their environment are able to survive and reproduce and goto next generation. In simple words, they simulate “survival of the fittest” among individual ofconsecutivegenerationforsolvingaproblem.Each generation consistof apopulation ofindividuals and each individual represents a point in search space and possible solution. Eachindividual is represented as a string of character/integer/float/bits. This stringis analogous totheChromosome.Geneticalgorithmsarebasedonananalogywithgeneticstructureandbehavior of chromosome of the population. Following is the foundation of GAs based on thisanalogy–

* + - Individualinpopulationcompeteforresourcesandmate
    - Thoseindividualswhoaresuccessful(fittest)thenmatetocreatemoreoffspringthanothers
    - Genes from “fittest” parent propagate throughout the generation, that is sometimesparentscreateoffspringwhichisbetterthaneitherparent.
    - Thuseachsuccessivegenerationis moresuitedfortheir environment.

The population of individuals are maintained within search space. Each individual representa solution in search space for given problem. Each individual is coded as a finite lengthvector(analogoustochromosome)ofcomponents.Thesevariablecomponentsare

analogoustoGenes. Thusachromosome(individual)iscomposedofseveralgenes(variablecomponents)

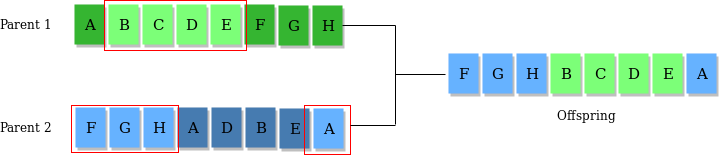


AFitnessScoreisgiventoeachindividualwhich**showstheabilityofanindividual to “compete”**. The individual having optimal fitness score (or near optimal) aresought.The GAs maintains the population of n individuals (chromosome/solutions) along withtheirfitnessscores.Theindividualshavingbetterfitnessscoresare givenmorechancetoreproduce than others. The individuals with better fitness scores are selected who mate andproduce **better offspring** by combining chromosomes of parents. The population size is staticso the room has to be created for new arrivals. So, some individuals die and get replacedbynew arrivals eventually creating new generation when all the mating opportunity of the oldpopulation is exhausted. It is hoped that over successive generations better solutions will arrivewhile least fit die.Each new generation has on average more “better genes” than the individual(solution)ofpreviousgenerations.Thuseachnewgenerationshavebetter**“partialsolutions”**than previousgenerations.Oncethe offspringsproducedhavingnosignificantdifference than offspring produced by previous populations, the population is converged. Thealgorithmissaidtobe convergedtoasetofsolutionsforthe problem.

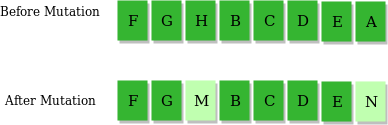
## OPERATORSOFGENETICALGORITHM:

Oncetheinitialgenerationiscreated,thealgorithmevolvethegenerationusingfollowingoperators–

1. **SelectionOperator:**Theideaistogivepreferencetothe individualswithgoodfitnessscores andallow themtopassthere genes tothe successive generations.
2. **Crossover Operator:** This represents mating between individuals. Two individuals areselected using selection operator and crossover sites are chosen randomly. Then the genes atthesecrossoversitesareexchangedthuscreatingacompletelynewindividual(offspring).Forexample–



**MutationOperator:**Thekeyideaistoinsertrandomgenesinoffspringtomaintainthediversityin populationtoavoid the prematureconvergence.Forexample–



## ALGORITHM:

1. Randomlyinitializepopulationsp
2. Determinefitnessofpopulation
3. Untillconvergencerepeat:
   1. Selectparentsfrompopulation
   2. Crossoverandgeneratenewpopulation
   3. Performmutationonnewpopulation
   4. Calculatefitnessfornewpopulation

## USESOFGENETICALGORITHM:

* + TheyareRobust
  + Provideoptimisationoverlargespacestate.
  + UnliketraditionalAI,theydonotbreakonslightchangeininputor presenceofnoise

**SOURCE CODE:**

importnumpydefcal\_pop\_fitness(equation\_inputs,pop):

#Calculatingthefitnessvalueofeachsolutioninthecurrentpopulation.

# The fitness function calulates the sum of products between each input and itscorrespondingweight.

fitness=numpy.sum(pop\*equation\_inputs,axis=1)returnfitness

defselect\_mating\_pool(pop,fitness,num\_parents):

#Selectingthebestindividualsinthecurrentgenerationasparentsforproducingtheoffspringofthenextgeneration.

parents=numpy.empty((num\_parents,pop.shape[1])) for parent\_num inrange(num\_parents):

max\_fitness\_idx=numpy.where(fitness==numpy.max(fitness)) max\_fitness\_idx =max\_fitness\_idx[0][0]parents[parent\_num,:]=pop[max\_fitness\_idx,:]fitness[max\_fitness\_idx]=-99999999999

returnparents

def crossover(parents,offspring\_size): offspring =numpy.empty(offspring\_size)

#Thepointatwhichcrossovertakesplacebetweentwoparents.Usually,itisatthecenter.crossover\_point= numpy.uint8(offspring\_size[1]/2)

forkinrange(offspring\_size[0]):

# Index of the first parent tomate. parent1\_idx =k%parents.shape[0]

# Index of the second parent to mate.parent2\_idx=(k+1)%parents.shape[0]

# The new offspring will have its first half of its genes taken from the firstparent.offspring[k,0:crossover\_point]=parents[parent1\_idx,0:crossover\_point]

# The new offspring will have its second half of its genes taken from the second parent.offspring[k,crossover\_point:]=parents[parent2\_idx,crossover\_point:]

returnoffspring

defmutation(offspring\_crossover,num\_mutations=1):

mutations\_counter=numpy.uint8(offspring\_crossover.shape[1]/num\_mutations)

# Mutation changes a number of genes as defined by the num\_mutationsargument. Thechanges are random.

foridxinrange(offspring\_crossover.shape[0]):gene\_idx=mutations\_counter-1formutation\_numinrange(num\_mutations):# The randomvalueto beaddedtothegene.

random\_value = numpy.random.uniform(-1.0, 1.0,1)

offspring\_crossover[idx, gene\_idx] = offspring\_crossover[idx,gene\_idx]+random\_value

gene\_idx = gene\_idx +mutations\_counter returnoffspring\_crossover

importnumpy

"""

They=targetistomaximize thisequationASAP:y=w1x1+w2x2+w3x3+w4x4+w5x5+6wx6where (x1,x2,x3,x4,x5,x6)=(4,-2,3.5,5,-11,-4.7)

Whatarethebestvaluesforthe6weightsw1tow6?

We are going to use the genetic algorithm for the best possible values after anumberofgenerations.

"""

#Inputsoftheequation.equation\_inputs = [4,-

2,3.5,5,-11,-4.7]

# Number of the weights we are looking tooptimize.num\_weights=len(equation\_inputs)

"""

Genetic algorithmparameters: Matingpool sizePopulationsize

"""

sol\_per\_pop=8

num\_parents\_mating=4

#Definingthepopulationsize.

pop\_size=(sol\_per\_pop,num\_weights)#Thepopulationwillhavesol\_per\_popchromosome where eachchromosome hasnum\_weights genes.

#Creatingtheinitialpopulation.

new\_population=numpy.random.uniform(low=-4.0,high=4.0,size=pop\_size)print(new\_population)

"""

new\_population[0,:]=[2.4,0.7,8,-2,5,1.1]

new\_population[1,:]=[-0.4,2.7,5,-1,7,0.1]

new\_population[2,:]=[-1,2,2,-3,2,0.9]

new\_population[3,:]=[4, 7,12,6.1,1.4,-4]

new\_population[4,:]=[3.1,4,0,2.4,4.8,0]

new\_population[5,:]=[-2,3,-

7,6, 3, 3]"""

best\_outputs = []num\_generations=1000

forgenerationinrange(num\_generations):

print("Generation : ",generation)

# Measuring the fitness of each chromosome in thepopulation.fitness=cal\_pop\_fitness(equation\_inputs,new\_population)print("Fitness")

print(fitness)best\_outputs.append(numpy.max(numpy.sum(new\_population\*equation\_inputs,

axis=1)))#Thebestresultinthecurrentiteration.

print("Bestresult:",numpy.max(numpy.sum(new\_population\*equation\_inputs,axis=1)))

#Selectingthebestparentsinthepopulationformating.parents =select\_mating\_pool(new\_population,fitness,num\_parents\_mating)

print("Parents")

print(parents)

#Generatingnextgenerationusing

crossover.offspring\_crossover=crossover(parents,

offspring\_size=(pop\_size[0]-parents.shape[0],num\_weights))

print("Crossover")print(offspring\_crossover)

#Addingsome variationstotheoffspringusing mutation.offspring\_mutation = mutation(offspring\_crossover,num\_mutations=2)print("Mutation")print(offspring\_mutation)

#Creating thenewpopulationbasedontheparentsandoffspring.new\_population[0:parents.shape[0],:]=parentsnew\_population[parents.shape[0]:,:]=offspring\_mutation

#Getting thebestsolutionafteriteratingfinishingallgenerations.#At first, the fitness is calculated for each solution in the finalgeneration.fitness= cal\_pop\_fitness(equation\_inputs,new\_population)

# Then return the index of that solution corresponding to thebest fitness. best\_match\_idx = numpy.where(fitness ==numpy.max(fitness))

print("Bestsolution:",

new\_population[best\_match\_idx, :]) print("Bestsolution fitness : ", fitness[best\_match\_idx])importmatplotlib.pyplotmatplotlib.pyplot.plot(best\_outp

uts)matplotlib.pyplot.xlabel("Iteration")matplotlib.pyplot.ylabel("Fitness")matplotlib.pyplot.show()

## OUTPUT:

[[0.582041412.32880696-2.951302092.570569533.33055238-0.58167871][-1.65052225

3.52263842-2.46577305-1.7005396-3.804802020.29677167] [ 2.6239874-2.01548549-

1.722922953.61090243-1.25604726-2.32647264][-3.451673932.857718253.74655682-

2.01790626 0.25750106-3.12923247][2.86026334-0.4306777-3.262979561.74863348-

1.93705571 -3.18855672][-1.700120890.98685104-1.911920723.91873942-0.09354385

1.43038667] [ 0.31769009 -0.87290809 3.75249785 2.57657993 0.588830822.83231871] [

3.833149260.33838112-2.49509594-1.507631743.99440509-0.03037715]]

Generation:

0

Fitness

[-33.708344139.6777259451.30214363-4.6238336545.91897711-1.566046069.24418172-

45.41084308]

Bestresult:

51.302143629097614

Parents

[[2.6239874-2.01548549-1.722922953.61090243-1.25604726-2.32647264][2.86026334-

0.4306777-3.262979561.74863348-1.93705571-3.18855672][-1.650522253.52263842-

2.46577305-1.7005396-3.804802020.29677167][0.31769009-0.872908093.75249785

2.576579930.588830822.83231871]]

Crossover

[[2.6239874-2.01548549-1.722922951.74863348-1.93705571-3.18855672][2.86026334-

0.4306777-3.26297956-1.7005396-3.804802020.29677167] [-1.650522253.52263842-

2.465773052.576579930.588830822.83231871][0.31769009-0.872908093.75249785

3.61090243-1.25604726-2.32647264]]

Mutation

[[2.6239874-2.01548549-1.678966321.74863348-1.93705571-3.97789372][2.86026334-

0.4306777-3.12878279-1.7005396-3.80480202-0.15430324][-1.650522253.52263842-

3.376696012.576579930.588830822.25153466][0.31769009-0.872908092.93428907

3.61090243-1.25604726-2.71597954]]

.

.

.

.

Generation :999

Fitness

[2554.39355622551.723607382549.405839542549.299316292552.24225166

2550.45506206

2547.12995122551.22467397]

Bestresult:

2554.3935561987346

Parents

[[ 3.17690088e-01 -8.72908094e-01 2.67689952e+02 1.74863348e+00 -1.93705571e+00-3.37108802e+02][3.17690088e-01 -8.72908094e-01

2.67638232e+021.74863348e+00-

1.93705571e+00-3.36689592e+02][3.17690088e-01-8.72908094e-01 2.67254110e+02

1.74863348e+00-1.93705571e+00-3.36865291e+02][3.17690088e-01-8.72908094e-01

2.67370854e+021.74863348e+00-1.93705571e+00-3.36672197e+02]]

Crossover

[[ 3.17690088e-01 -8.72908094e-01 2.67689952e+02 1.74863348e+00 -1.93705571e+00-3.36689592e+02]

[3.17690088e-01-8.72908094e-012.67638232e+021.74863348e+00

-1.93705571e+00-3.36865291e+02]

[3.17690088e-01-8.72908094e-012.67254110e+021.74863348e+00

-1.93705571e+00-3.36672197e+02]

[ 3.17690088e-01 -8.72908094e-012.67370854e+021.74863348e+00

-1.93705571e+00-3.37108802e+02]]

Mutation

[[3.17690088e-01-8.72908094e-012.68382875e+021.74863348e+00

-1.93705571e+00-3.36222272e+02]

[3.17690088e-01-8.72908094e-012.68456819e+021.74863348e+00

-1.93705571e+00-3.37417363e+02]

[ 3.17690088e-01 -8.72908094e-012.67606746e+021.74863348e+00

-1.93705571e+00-3.36866918e+02]

[ 3.17690088e-01 -8.72908094e-012.67051753e+021.74863348e+00

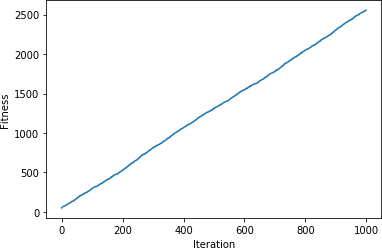
-1.93705571e+00-3.37331663e+02]]

Bestsolution:[[[3.17690088e-01-8.72908094e-01 2.68456819e+021.74863348e+00

-1.93705571e+00-3.37417363e+02]]]

Best solutionfitness

:[2558.52782726]



## VIVAQUESTIONS & ANSWERS

***1.*WhatisaGeneticAlgorithm?**

Let’sgetbacktotheexamplewediscussedaboveandsummarizewhatwedid.

1. Firstly,wedefinedour initialpopulationasourcountrymen.
2. Wedefinedafunctionto classifywhetherisapersonisgoodorbad.
3. Thenweselectedgoodpeopleformatingtoproducetheiroff-springs.
4. Andfinally,theseoff-springsreplacethebadpeoplefromthepopulationandthisprocessrepeats.

# PROGRAM 9

## IMPLEMENT THE FINITE WORDS CLASSIFICATION SYSTEM USING BACK-PROPAGATIONALGORITHM

**AIM:**

Toimplement thefinitewordsclassificationsystemusingBack-propagationalgorithm

## DESCRIPTION:

**Whatisbackpropagation?**

We can define the backpropagation algorithm as an algorithm that trains some given feed-forward Neural Network for a given inputpattern wherethe classifications are known tous.At the point when every passage of the example set is exhibited to the network, the networklooks atits yield reaction to the exampleinputpattern. After that,thecomparison donebetween output response and expected output with the error value is measured.Later,weadjusttheconnectionweightbasedupontheerrorvaluemeasured.

It was first introduced in the 1960s and 30 years later it was popularized by DavidRumelhart, Geoffrey Hinton, and Ronald Williamsin the famous 1986 paper. In this paper,they spoke about the various neural networks. Today, back propagation is doing good. Neuralnetworktraininghappensthroughbackpropagation.Bythisapproach,wefine-tunetheweights of a neural net based on the errorrate obtained in the previous run. The rightmannerofapplyingthistechniquereduceserrorratesandmakesthemodelmorereliable.Backpropagation is used to train theneural network of thechain rulemethod.In simpleterms,aftereachfeed-forwardpassesthroughanetwork,thisalgorithmdoesthebackward

pass to adjustthe model’s parameters based on weights andbiases.A typical supervisedlearning algorithm attempts to find a function that maps input data to the right output. Backpropagation works with a multi-layered neural network and learns internal representations ofinputtooutputmapping.

The Back propagation algorithm is a supervised learning method for multilayer feed-forwardnetworksfromthefieldofArtificialNeuralNetworks.

Feed-forward neural networks are inspired by the information processing of one or more neuralcells, called a neuron. A neuron accepts input signals via its dendrites, which pass the electricalsignal down to the cell body.The axon carriesthe signal out tosynapses,which are theconnections ofacell’saxontoothercell’s dendrites.

Theprincipleofthebackpropagationapproachistomodelagivenfunctionbymodifyinginternal weightings of inputsignals to produce an expected output signal.Thesystem is trained using a supervised learning method, where the error between the system’soutput and a known expected output is presented to the system and used to modify its internalstate.

Technically, the backpropagation algorithm is a method for training the weights in a multilayerfeed-forward neural network. As such, it requires a network structure to be defined of one ormore layers where one layeris fully connected to the nextlayer. A standard network structureis one input layer, one hidden layer, and one outputlayer.Back propagation can be used forbothclassificationandregressionproblems.

In classification problems, best results are achieved when the network has one neuron in theoutput layer for each class value. For example, a 2-class or binary classification problem withthe class values of A and B. These expected outputs would have to be transformed into binaryvectors with one column for each class value. Such as [1, 0] and [0, 1] for A and B respectively.Thisis calleda onehotencoding.

## Howdoesbackpropagationwork?

Letustakealookathowbackpropagationworks.Ithasfourlayers:inputlayer,hidden layer,hiddenlayerIIandfinaloutputlayer.

So,themainthreelayersare:

1. Inputlayer
2. Hiddenlayer
3. Outputlayer

Each layer has its own way of working and its own way to take action such that we are able toget the desired results and correlate these scenarios to our conditions. Let us discuss otherdetailsneededtohelpsummarizingthisalgorithm.

## SOURCECODE:

importpandasaspd

fromsklearn.model\_selectionimport train\_test\_splitfromsklearn.feature\_extraction.textimportCountVectorizer from sklearn.neural\_networkimportMLPClassifier

fromsklearn.metricsimportaccuracy\_score,confusion\_matrix,precision\_score,recall\_score

msg= pd.read\_csv('document.csv',names=['message', 'label']) print("Total Instances ofDataset:",msg.shape[0])msg['labelnum']=msg.label.map({'pos':1,'neg':0})

X =

msg.messagey

=

msg.labelnum

Xtrain,Xtest,ytrain,ytest=train\_test\_split(X,y)

count\_v = CountVectorizer()Xtrain\_dm =count\_v.fit\_transform(Xtrain)Xtest\_dm =count\_v.transform(Xtest)

df=pd.DataFrame(Xtrain\_dm.toarray(),columns=count\_v.get\_feature\_names())

clf = MLPClassifier(solver='lbfgs', alpha=1e-5,hidden\_layer\_sizes=(5, 2), random\_state=1)clf.fit(Xtrain\_dm,ytrain)pred= clf.predict(Xtest\_dm)

print('AccuracyMetrics:')

print('Accuracy:',accuracy\_score(ytest,pred))print('Recall:',recall\_score(ytest,pred))print('Precision: ', precision\_score(ytest,pred))

print('ConfusionMatrix:\n',confusion\_matrix(ytest,pred))

## document.csv:

Ilove thissandwich,posThisis anamazingplace,pos

I feel very good about thesebeers,pos This is my bestwork,pos

What an awesome view,posIdonotlikethisrestaurant,negIam

tired of this stuff,negIcan'tdealwiththis,neg Heis

my swornenemy,neg Mybossishorrible,neg

This is an awesome place,posIdonotlikethe tasteof

thisjuice,negIlovetodance,pos

Iamsickandtiredofthisplace,negWhatagreatholiday,pos

Thatisabadlocalitytostay,negWe willhavegoodfun

tomorrow,posIwenttomy

enemy'shousetoday,neg

## OUTPUT:

Total Instances ofDataset: 18AccuracyMetrics:

Accuracy:0.8

Recall: 1.0Precisio

n: 0.75

ConfusionMatrix:

[[1 1]

[03]

## VIVAQUESTIONS&ANSWERS:

1. **WhatisthedifferencebetweenaPerceptronandLogisticRegression?**

A Multi-Layer Perception (MLP) is one of the most basic [neuralnetworks](https://courses.analyticsvidhya.com/courses/Introduction-to-Neural-Networks?utm_source=blog&utm_medium=comprehensive-popular-deep-learning-interview-questions-answers)that we use forclassification.For abinaryclassificationproblem,weknowthattheoutputcanbeeither0or1.This is just like our simple logistic regression, where we use a logit function to generate aprobabilitybetween0and1.

So,what’sthe differencebetweenthetwo?

Simplyput,itisjustthedifferenceinthethresholdfunction!Whenwerestrictthelogisticregressionmodeltogiveuseitherexactly1orexactly0,wegetaPerceptionmodel:



## Canwehavethesamebiasforallneuronsofa hiddenlayer?

Essentially, youcanhaveadifferentbiasvalueateachlayerorateach neuron aswell.However,itisbestifwehaveabiasmatrixforalltheneuronsinthehiddenlayersaswell.

Apointtonoteisthatboththesestrategieswouldgiveyouverydifferentresults.

## Whatifwedonotuse anyactivationfunction(s)ina neuralnetwork?

Themainaimofthisquestionistounderstandwhyweneed[activationfunctions](https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/?utm_source=blog&utm_medium=comprehensive-popular-deep-learning-interview-questions-answers)inaneuralnetwork.Youcanstartoffbygivingasimpleexplanationofhowneuralnetworksarebuilt:

**Step1:**Calculatethesumofalltheinputs(X)accordingtotheirweightsandincludethebiasterm:

Z=(weights\*X)+bias

**Step 2:**Applyanactivationfunctiontocalculatetheexpectedoutput:Y =Activation(Z)

Steps1and 2 areperformed at each layer. Ifyou recollect,this isnothing but forwardpropagation!Now,whatifthereisnoactivationfunction?

OurequationforYessentiallybecomes:

Y=Z=(weights \* X)+bias

Wait – isn’t this just a simple linear equation? Yes – and that is why we need activationfunctions. A linear equation will not be able to capture the complex patterns in the data – this isevenmoreevidentinthecaseofdeeplearningproblems.

In order to capture non-linear relationships, we use activation functions, and thatis why aneuralnetworkwithoutanactivationfunctionisjustalinearregression model.

1. **Inaneuralnetwork,whatifalltheweightsareinitializedwiththesamevalue?**

In simplest terms, if all the neurons have the same value of weights, each hidden unit will getexactly the same signal. While this might work during forward propagation, the derivative ofthe costfunctionduringbackwardpropagationwouldbe the same everytime.

In short, there is no learning happening by the network! What do you call the phenomenon ofthemodelbeingunabletolearnanypatternsfromthedata?Yes,[underfitting.](https://www.analyticsvidhya.com/blog/2020/02/underfitting-overfitting-best-fitting-machine-learning/?utm_source=blog&utm_medium=comprehensive-popular-deep-learning-interview-questions-answers)

Therefore,ifallweightshavethesameinitialvalue,thiswouldleadtounderfitting.

1. **Whatis theroleofweights andbias inaneuralnetwork?**

This is a question best explained with a real-life example. Consider that you want to go outtoday to play a cricket match with your friends. Now, a number of factors can affect yourdecision-making,like:

* + Howmanyofyourfriendscanmakeittothegame?
  + Howmuchequipmentcanallofyoubring?
  + Whatisthetemperatureoutside?

And so on. These factors can change your decision greatly or not too much. For example, if it israining outside, then you cannot go out to play at all. Or if you have only one bat, you can shareit while playing as well. The magnitude by which these factors can affect the game is called theweightofthatfactor.

Factors like the weather or temperature might have a higher weight, and other factors likeequipmentwouldhave alowerweight.

However, does this mean that we can play a cricket match with only one bat? No – we wouldneed 1 ball and 6 wickets as well. This is where bias comes into the picture. Bias letsyouassign some threshold which helps you activate a decision-point (or a neuron) only when thatthresholdis crossed.